Lending Club Case Study

Seyed Hamid Reza Hosseini

Outline

- Problem statement
- Data Understanding
- Data Cleaning
- Data Analysis
- Recommendations
- Acknowledgements

Problem statement

It is important for a Lending Club to identify which group of customers might default, meaning they might be charged off and don't return the money, based on their application. If this identification is not performed correctly, it can lead to loss of business in two ways:

- If the application of the customers who fully pay their loan, is mistakenly rejected, then this will lead to the loss of business.
- If the application of the defaulters are approved, and then later they are charged off and don't return the money back, then this will also lead to loss of business.

In this project the dataset, containing the history of customers of a Lending Club is given, with the information about whether each customer has been charged off or fully paid back the loan.

The objective is to identify the groups of customers who are most likely to default, and advise this to the business.

Data Understanding

- The dataset of the Lending Club was imported.
- The columns with more than 50% of NaN (null) values were removed, and hence the number of columns dropped to 54 from 111.
- The following columns were also removed:
    - Customer behaviour: since these are used at the time of revieweing the loan application and before approving the application. However, now that all the applications of the customers are approved, these columns are irrelevant to this study
    - the geographic location of the applicant
    - columns that have: (i) only one unique value; (ii) IDs; (iii) only 0 or NaN; (iv) ‘sub_grade’, due to no need for further granularity; (v) URL, description, loan title and employment title since these have no impact on the loan_status.
- The rows that have ‘Current’ value in loan_status column were removed from the dataframe, since it is unknown whether they will be charged off (default) or not.

Data Cleaning

- The columns 'emp_length' and 'pub_rec_bankruptcies’ are both categorical and still have null values. For these columns, the most frequent element in the column is replaced with the NaN values. (the number of rows (data entries) that are affected are less than 4.50% of the entire rows. Hence, this doesn't significantly impact the analysis.)
- Sanity check performed to ensure:

  funded_amnt_inv <= funded_amnt <= loan_amnt

Data Cleaning (contd.)

- Preparing columns for data analysis
  - The ‘loan_status’ with the entries of 'Fully Paid', 'Charged Off’ were transformed to ‘0’ and ‘1’, respectively, in a new column called ‘charged_off?’.
  - The ‘%’ at the end of the elements in interest rate were removed.
  - All the columns were categorised into bins in order to analyse the data in manageable number of categories rather than in numeric values or in higher number of categories.

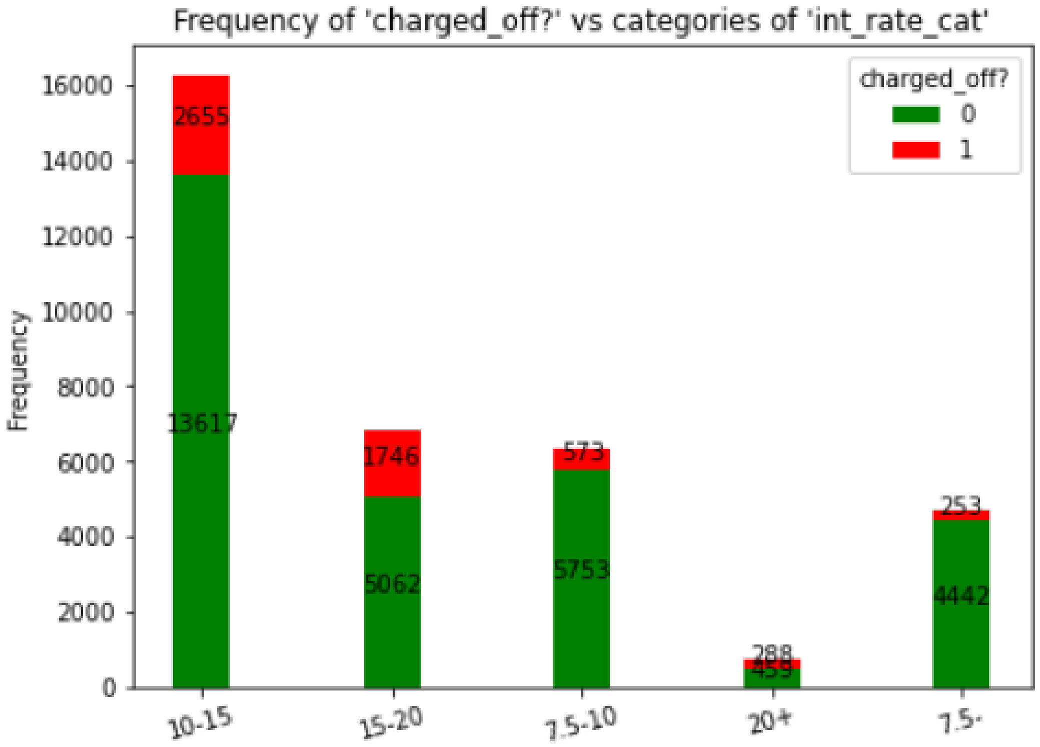
In this way, the data were prepared for the next step, i.e. data analysis.

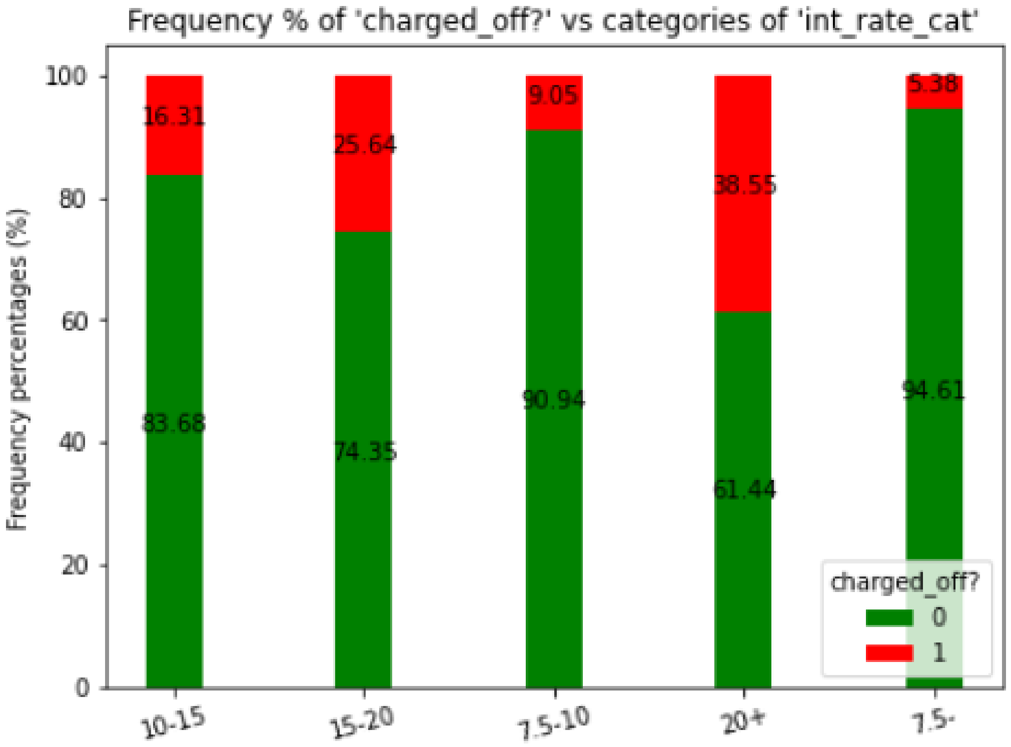
Data Analysis

- The new categorised columns were gathered into a new dataframe, however, this new dataframe has several duplicates due to the new categories. Hence, the duplicates were removed and the number of entries (rows) reduced to 34848 from 38577.
- For every column of this dataframe, two plots were produced:
  - frequency count plots of categories of ‘charged_off?’ over the new categories of the column
  - percentage frequency count plots of the above plot: in order to observed the share (percentage) of defaulters of every new category of the column.
- In the next slides, the most important drivers of the defaulters are presented.

Data Analysis (contd.)

Impact of 'int_rate_cat' on 'charged_off?'

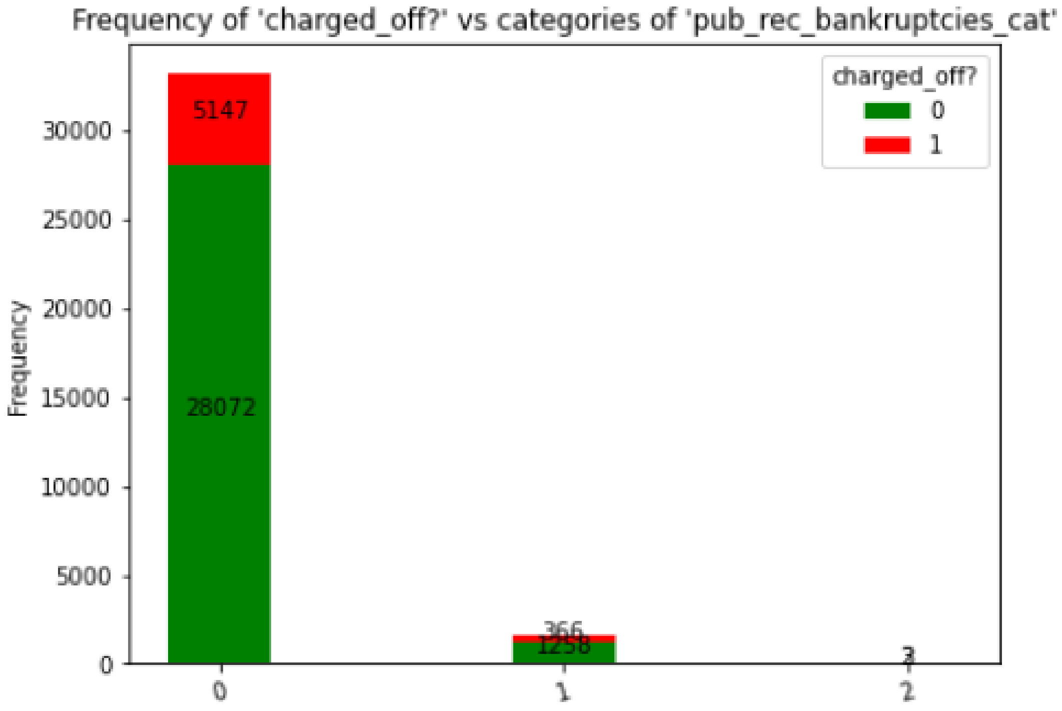




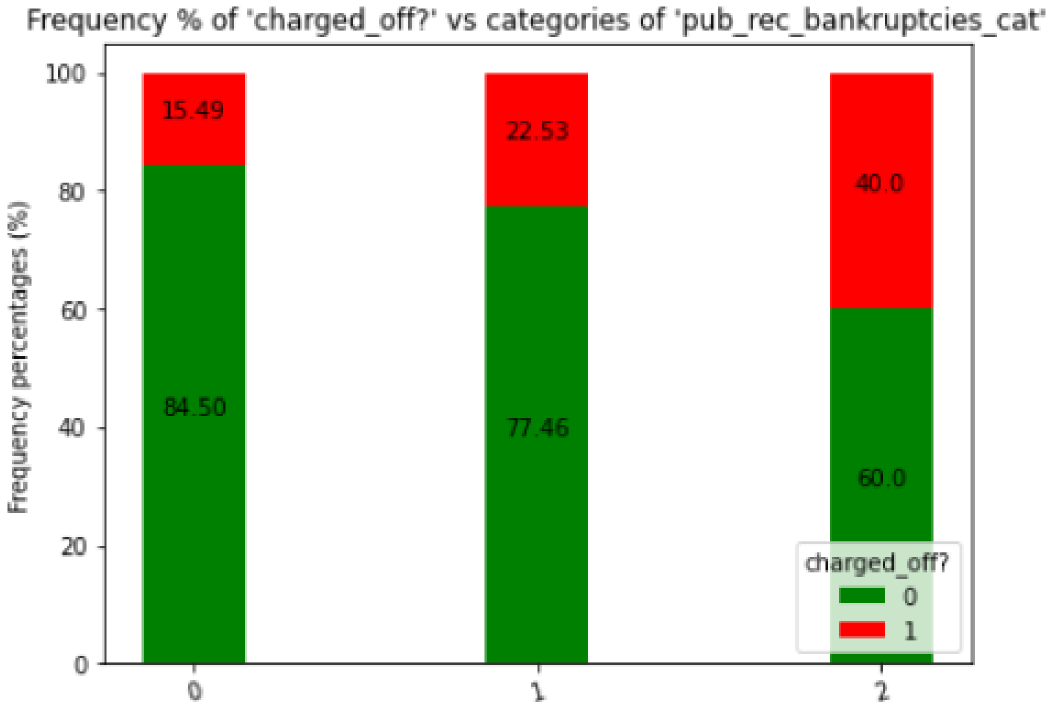
As can be seen, the **interest rate of more than 20%** can potentially be considered as one of the drivers of the customer being charged off, since the percentage of the defaulters in this category of interest rate is much higher than the same percentage in the rest of the categories.

Data Analysis (contd.)

Impact of 'pub_rec_bankruptcies_cat' on 'charged_off?'



Frequency of 'charged_off?' vs categories of 'pub_rec_bankruptcies_cat'

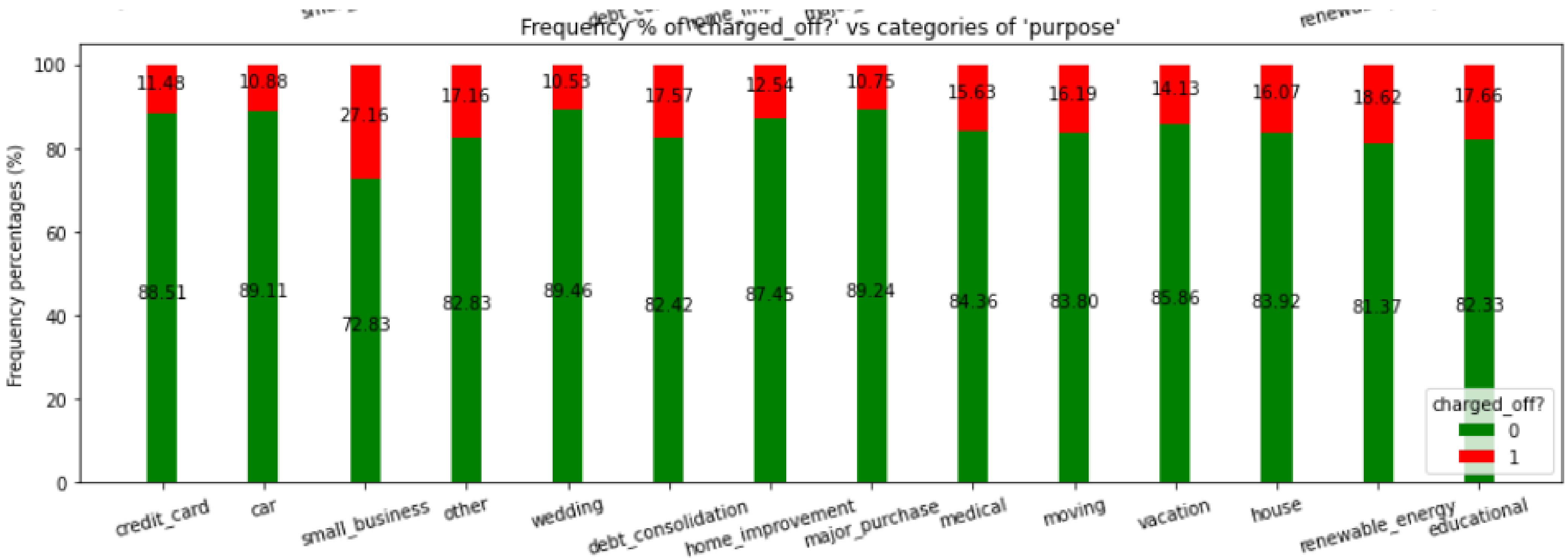


Frequency % of 'charged_off?' vs categories of 'pub_rec_bankruptcies_cat'

The customers with **2 public records of bankruptcies** are very probable to default.

Data Analysis (contd.)

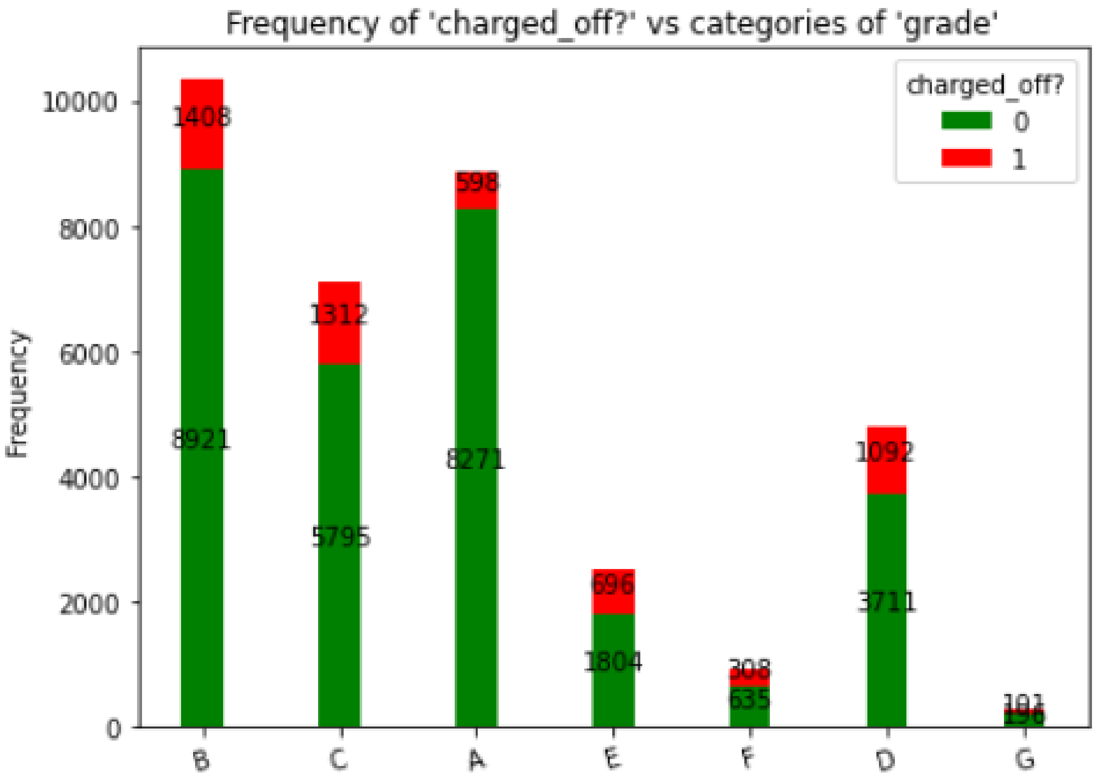
Impact of ‘purpose' on 'charged_off?'

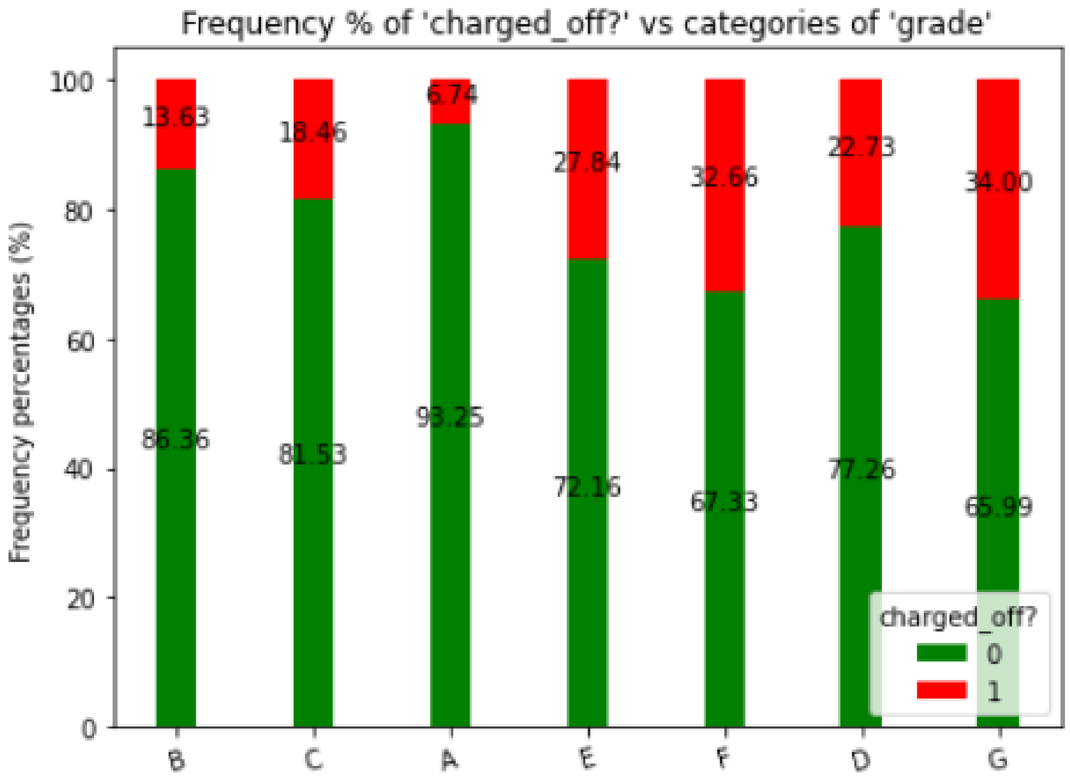


The **loan purpose of 'small business'** can also be considered as another driver for the customer to default.

Data Analysis (contd.)

Impact of ‘grade' on 'charged_off?'

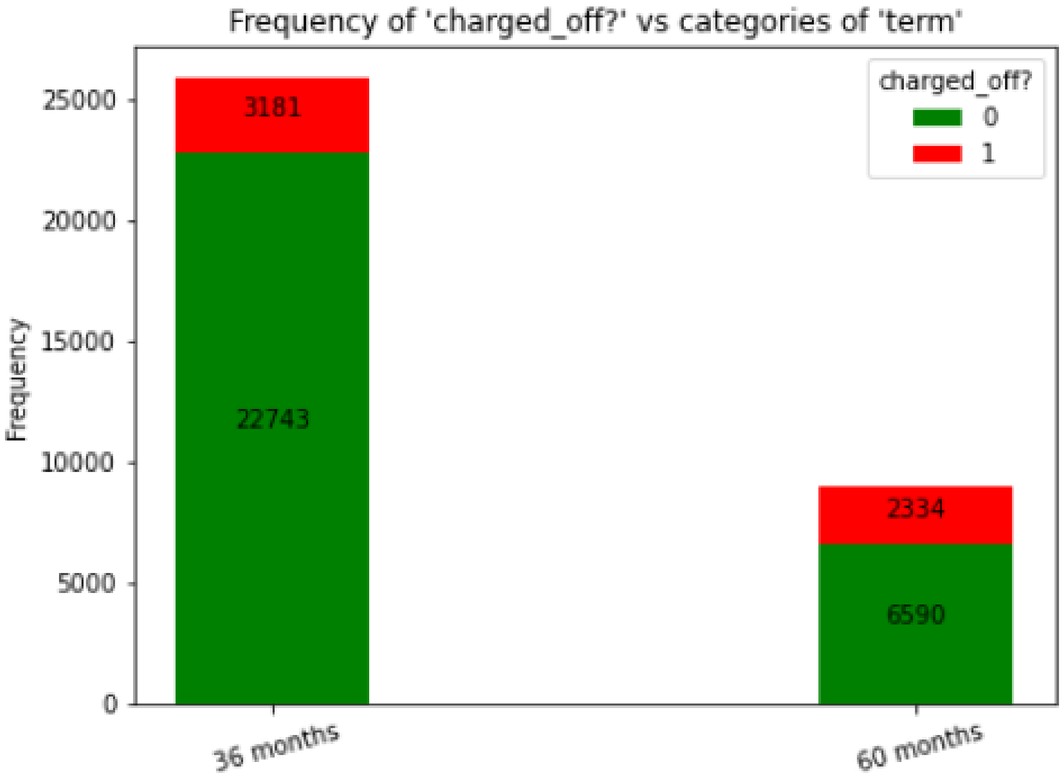


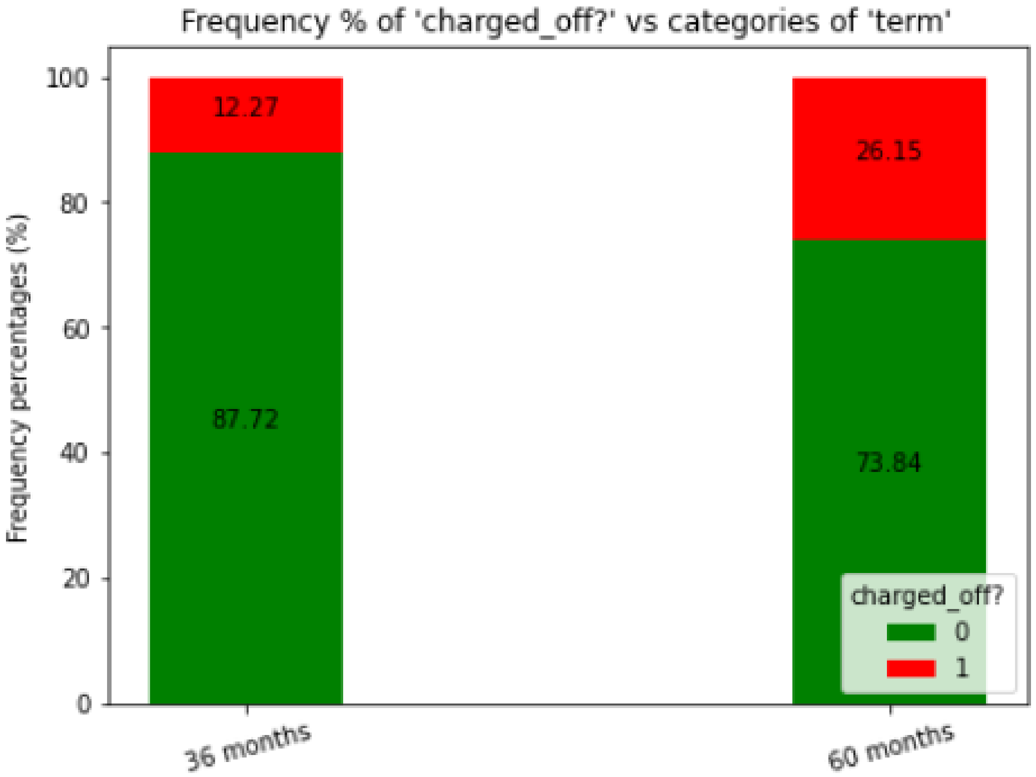


As can be seen, customers with **<u>LC assigned loan grade of 'G' </u>**is very probable to default due to the much higher percentage of defaulters in this category.

Data Analysis (contd.)

Impact of ‘term' on 'charged_off?'

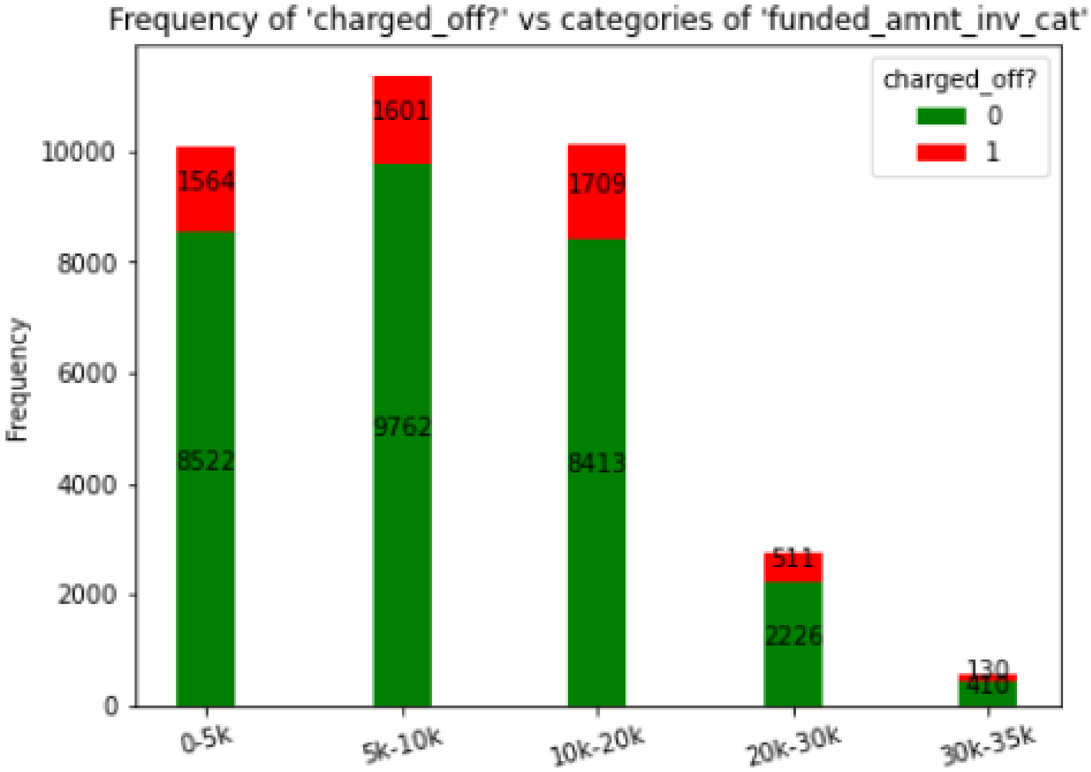




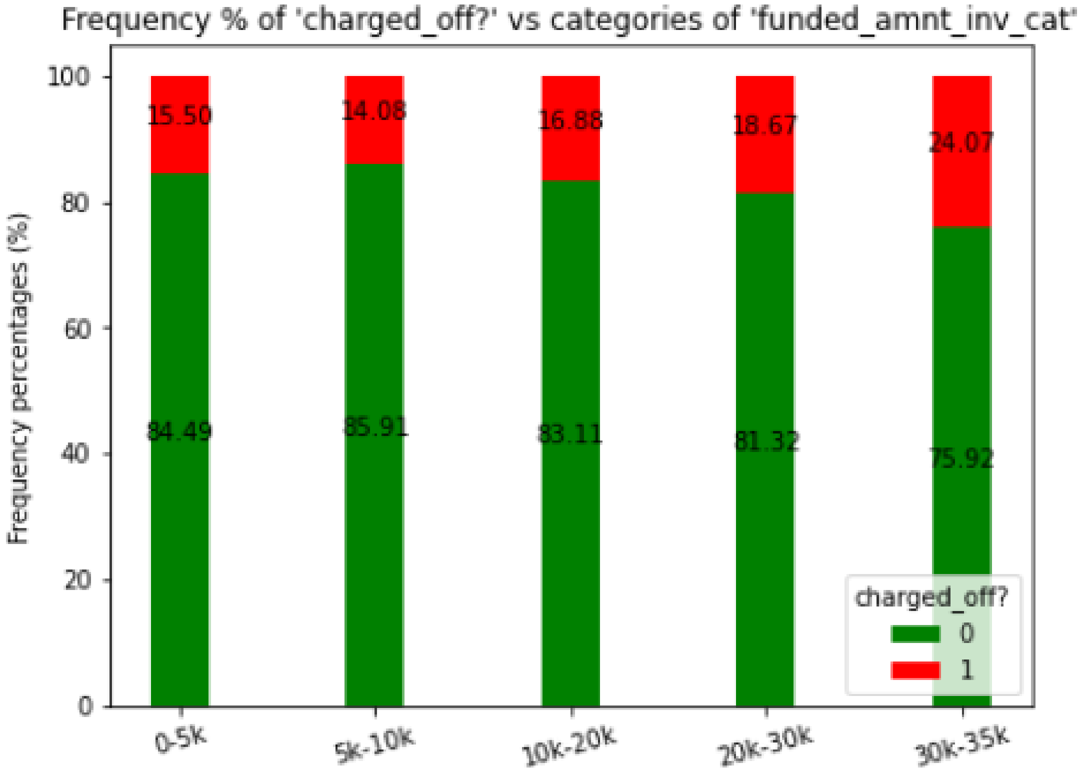
The loans with **<u>60 months return term</u>** is very probable to be charged off. Hence, this category of loan term can be considered as one of the drivers to default.

Data Analysis (contd.)

Impact of 'funded_amnt_inv_cat', 'funded_amnt_cat', and 'loan_amnt_cat' on 'charged_off?'



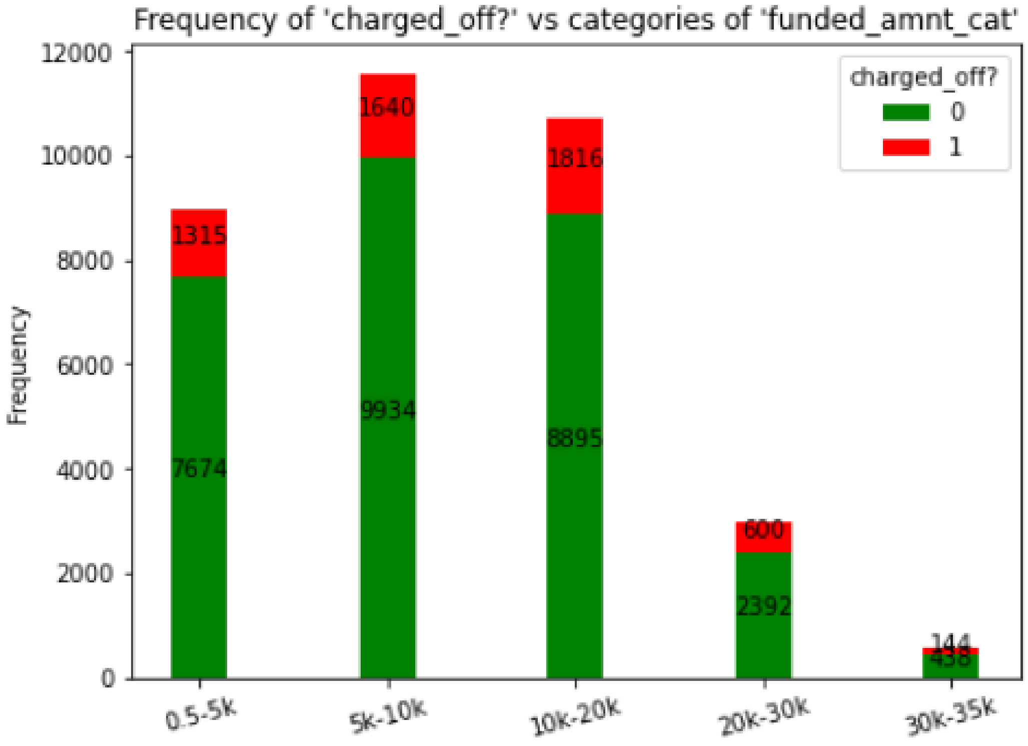
Frequency of 'charged_off?' vs categories of 'funded_amnt_inv_cat'

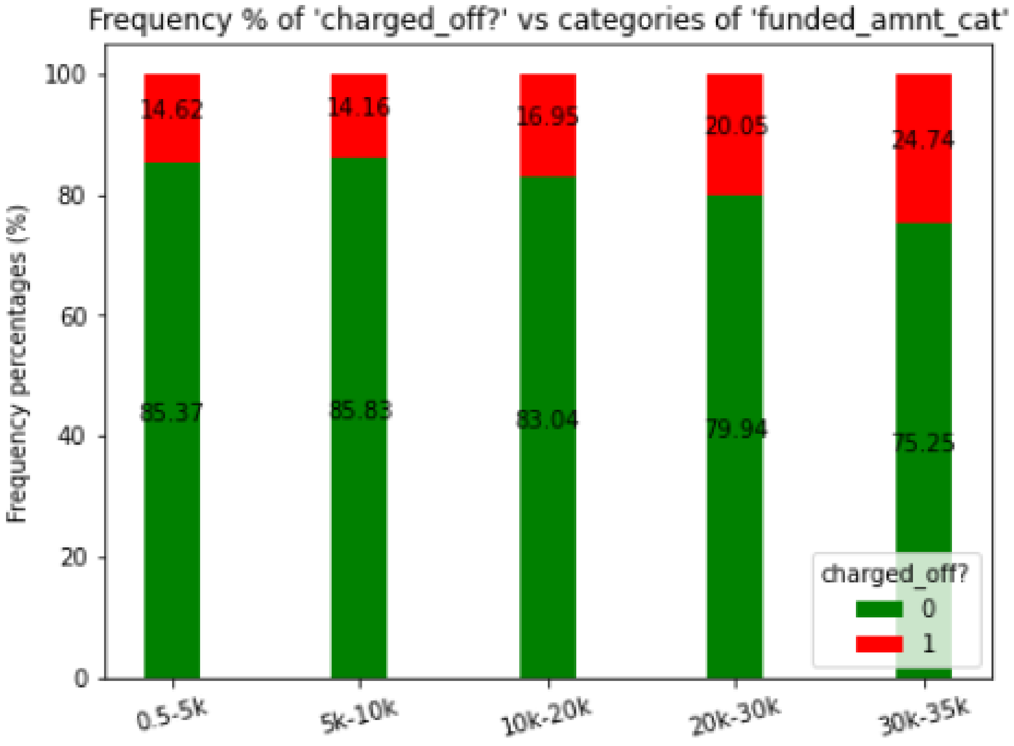


Frequency % of 'charged_off?' vs categories of 'funded_amnt_inv_cat'

Data Analysis (contd.)

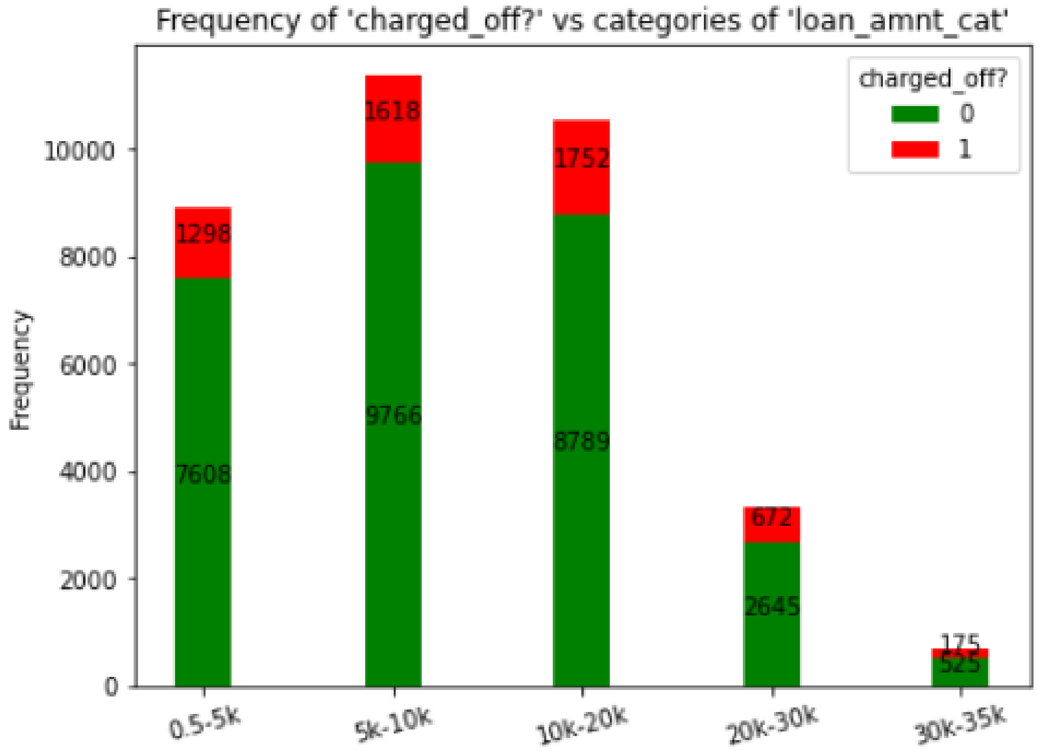
Impact of 'funded_amnt_inv_cat', 'funded_amnt_cat', and 'loan_amnt_cat' on 'charged_off?' (contd.)

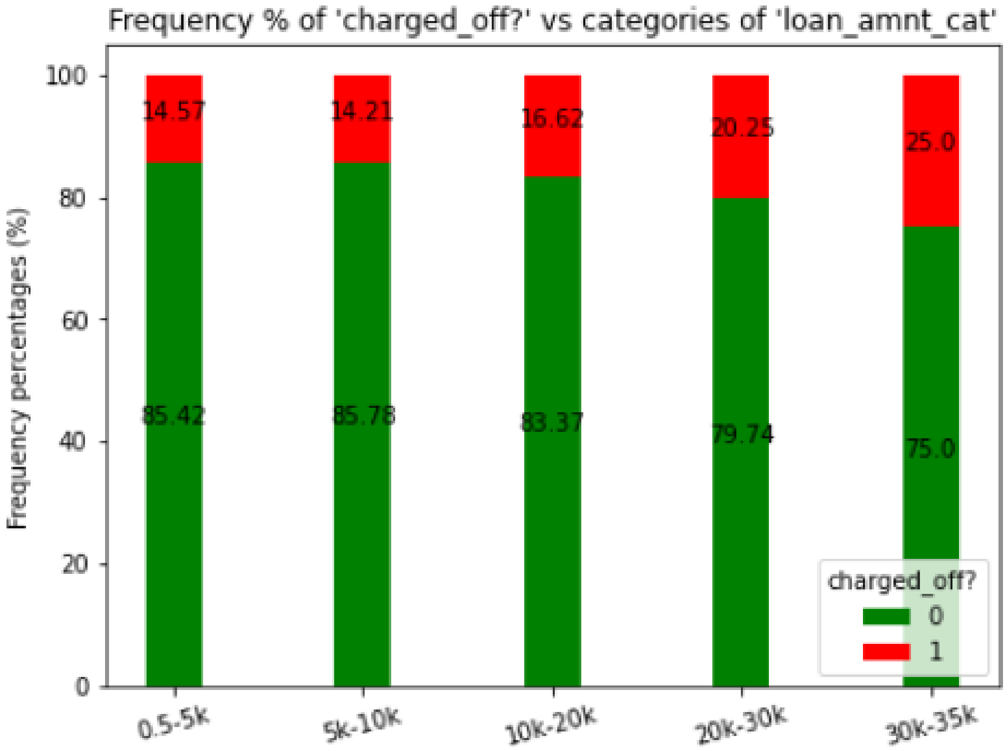




Data Analysis (contd.)

Impact of 'funded_amnt_inv_cat', 'funded_amnt_cat', and 'loan_amnt_cat' on 'charged_off?' (contd.)





As can be seen, in all the above graphs, the percentage share of defaulters in the **'30k-35k' category** is higher compared to the rest of the categories. Hence, this category for 'funded_amnt_inv_cat', 'funded_amnt_cat', and 'loan_amnt_cat' can be considered as one of the drivers for the customer to be charged off.

Recommendations

Analysis of the data of the Lending Club reveals that the customers within the following categories are most probable to default and cause loss of business to the Lending Club:

- interest rate more than 20%
- 2 public records of bankruptcy
- loan term of 60 months
- loan of 30k-35k during all the stages of application by the customer, approving by the club, and investing by the investors.
- loan purpose of 'small business'
- LC assigned loan grade of 'G'

Acknowledgements

- I would like to acknowledge the feedback, support and dataset provision by upGrad and The International Institute of Information Technology (IIIT), Bangalore.
- Also, I would like to express my gratitude to Aditya Bhattacharya for providing clarification and guidance to carry out this project.
- Furthermore, the valuable feedback from Dr Tayeb Jamali is highly appreciated.